StyleSwin: Transformer-based GAN for High-resolution Image Generation

Year: 2021 | Citations: 284 | Authors: Bo Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen

---

Abstract

Despite the tantalizing success in a broad of vision tasks, transformers have not yet demonstrated on-par ability as ConvNets in high-resolution image generative modeling. In this paper, we seek to explore using pure transformers to build a generative adversarial network for high-resolution image synthesis. To this end, we believe that local attention is crucial to strike the balance between computational efficiency and modeling capacity. Hence, the proposed generator adopts Swin transformer in a style-based architecture. To achieve a larger receptive field, we propose double attention which simultaneously leverages the context of the local and the shifted windows, leading to improved generation quality. Moreover, we show that offering the knowledge of the absolute position that has been lost in window-based transformers greatly benefits the generation quality. The proposed StyleSwin is scalable to high resolutions, with both the coarse geometry and fine structures benefit from the strong expressivity of transformers. However, blocking artifacts occur during high-resolution synthesis because performing the local attention in a block-wise manner may break the spatial coherency. To solve this, we empirically investigate various solutions, among which we find that employing a wavelet discriminator to examine the spectral discrepancy effectively suppresses the artifacts. Extensive experiments show the superiority over prior transformer-based GANs, especially on high resolutions, e.g., $1024 \times$ 1024. The StyleSwin, without complex training strategies, excels over StyleGAN on CelebA-HQ 1024, and achieves on-par performance on FFHQ-1024, proving the promise of using transformers for high-resolution image generation. The code and pretrained models are available at https://github.com/microsoft/StyleSwin.